Computer Physics Communications 269 (2021) 108126

Contents lists available at ScienceDirect

Computer Physics Communications

www.elsevier.com/locate/cpc

DynDen: Assessing convergence of molecular dynamics simulations of interfaces ☆,☆☆

Matteo T. Degiacomi [a,*], Shansi Tian [b], H. Chris Greenwell [c], Valentina Erastova [d,*]

[a] Department of Physics, Durham University, South Road, Durham, DH1 3LE, UK
[b] Key Laboratory of continental shale hydrocarbon accumulation and efficient development (Northeast Petroleum University), Ministry of Education, Northeast Petroleum University, Daqing, 163318, Heilongjiang, PR China
[c] Department of Earth Sciences, Durham University, South Road, Durham, DH1 3LE, UK
[d] School of Chemistry, Centre of Science at Extreme Conditions and the UK Centre of Astrobiology, University of Edinburgh, David Brewster Road, Edinburgh, EH9 3FJ, UK

ARTICLE INFO

Article history:
Received 7 December 2020
Received in revised form 13 July 2021
Accepted 22 July 2021
Available online 5 August 2021

Keywords:
Molecular dynamics
Interfaces
Layered materials
Equilibration
Convergence
Linear density

ABSTRACT

Molecular dynamics is a simulation technique used to predict the physical properties of systems based on their chemical structure and evolution of their atomic constituents. For these predictions to be reliable, it is critical that the simulation has reached convergence, whereby representative sampling of the phase space has been gathered. We show that the commonly used root mean square deviation is an unsuitable convergence descriptor for systems featuring surfaces and interfaces. We then present an effective criterion, embodied in the analysis tool *DynDen*, based on convergence of the linear partial density of all components in the simulation. With a variety of examples we demonstrate the usage of *DynDen* for the assessment of convergence, as well as for identification of slow dynamical processes, which can be easily missed with conventional analysis.

Program summary
Program title: DynDen
CPC Library link to program files: https://doi.org/10.17632/dfb34rpyw2.1
Developer's repository link: https://github.com/punkpony/DynDen
Licensing provisions: GNU General Public License 3
Programming Language: Python 3.X
External Routines/Libraries: MDAnalysis, numpy, matplotlib
Nature of problem: assessment of convergence in molecular dynamics simulations of interfaces
Solution method: reporting on the convergence of correlation of the linear density profiles of each individual system component

1. Introduction

Molecular dynamics (MD) is a computational technique aimed at inferring the physical and chemical properties of a molecular system by investigating its atomic structure and associated dynamics. Insights obtained by MD have been shown to be not only capable of explaining experimental data, but also to be sufficiently accurate to act as a predictive tool informing the design of new experiments. For this, the modelled system has to be representative of the macroscopic state. While this would often be regarded as system being *equilibrated*, we note that this does not mean that the system is in a complete thermodynamic equilibrium, neither in an experiment nor simulation. For instance, organic components in an oil reservoir are constantly undergoing slow evolution on geological timescales. In such case, both experiments and models probe the current state of the reservoir.

Through MD simulation the researcher can obtain an ensemble average of experimental observables through a time-average of the modelled dynamics. According to the ergodic theorem, an infinitely long simulation will be representative of the experimentally observed phenomenon. Yet, although improvements in both computing hardware and algorithmic efficiency are steadily ex-

☆ The review of this paper was arranged by Prof. Hazel Andrew.
☆☆ This paper and its associated computer program are available via the Computer Physics Communications homepage on ScienceDirect (http://www.sciencedirect.com/science/journal/00104655).

* Corresponding authors.
E-mail addresses: matteo.t.degiacomi@durham.ac.uk (M.T. Degiacomi), valentina.erastova@ed.ac.uk (V. Erastova).

https://doi.org/10.1016/j.cpc.2021.108126

tending the timescales reachable by simulations, ergodicity can never be assumed [1–4]. Often, simulation equilibrium is identified by a convergence in the system's root mean square deviation (RMSD) with respect of the starting conformation [5]. However, RMSD can be ill-suited to characterize systems featuring bulk solvents. Indeed, mixing of solvent molecules in the bulk phase, while not changing the overall properties of the system, will be reflected by an increasing RMSD. The highest possible RMSD in a simulation box will only be reached when every molecule has moved from its initial position by a distance proportional to the dimensions of the box itself, and therefore has explored all the available phase-space. Furthermore, RMSD is a global system property and, while simulations of interfaces often incorporate bulk, researchers are usually interested in assessing the local interactions of near-interface regions, at a defect or a specific surface site. In these regions, dynamics are different from those in the bulk, as they are restricted and, thus, slower.

In this article, we present a method and associated software, *DynDen*, to characterize the dynamics and to assess the convergence of simulations featuring interfaces. Using simulation data from our previous work on interfaces [6–8], we show that RMSD is insufficient to determine whether the simulation is biased by undersampling, leading to erroneous interpretations. The *DynDen* method is based on assessing the evolution of linear density profiles (the distribution of individual components as a function of distance orthogonal to interfacial region) and relates system convergence to the pairwise correlation coefficient of all simulation time frames [9]. Importantly, while a position swap of two identical molecules is associated with a change in RMSD, the linear density remains unaffected. However, when a position swap of two molecules of different type takes place, linear density will report a change, indicating a need for further sampling. Therefore, linear density is a suitable metric to identify (de)mixing while being insensitive to molecular shuffling in the bulk. When the linear density of each component in the system no longer evolves, one may assume the system to be kinetically converged.

2. Method

DynDen analyzes the evolution of linear densities of individual groups of atoms during the simulation. These groups can be either molecules of a given type, or any other user-defined atom selections. The linear densities are calculated orthogonal to the xy-plane of the interface. A good practice for all trajectory analysis, including this one, is to wrap each frame of the trajectory so that layers remain either in the centre of the box or at z=0. The *DynDen* pipeline is summarized as follows:

1. **Import trajectory and topology, reading user-defined parameters**. All file formats supported by the *Python MDAnalysis* [10,11] package are supported.
2. **Identify all the individual components of the system**. These are either automatically identified based on their residue name or declared by the user.
3. **Measure the length of the z-axis for each frame of the simulation** (Fig. 1B). The z-axis of each frame is subdivided in a parameter-defined number of bins to enable comparisons between different simulation frames. In the *NVT*-ensemble simulations bins will have the same size throughout the simulation, while for the *NPT*-ensemble simulations these may differ. This makes the analysis less reliable for cases where the bin size fluctuates beyond the sub-Ångstrom scale. Nevertheless, such large volume fluctuations are likely indicative of a non-converged system.
4. **Calculate linear density time series**. The linear density of a component is a histogram reporting on the number of its atoms along the binned z-axis. *DynDen* aggregates the linear densities of each simulation frame in a density time series plot (DTS, see example on the top row of Fig. 1D), representing the evolution of the linear density through the duration of the trajectory. The user has the option of averaging linear densities over multiple frames.
5. **Calculate pairwise correlation of density time series**. For each component, an all-vs-all correlation coefficient plot of the DTS is reported (pairwise density correlation, PDC, see example on the bottom row of Fig. 1D). Each coordinate (i, j) in the plot is the Pearson correlation coefficient $(\rho_{i,j})$ between linear densities at different simulation frames (d_i, d_j):

$$\rho_{i,j} = \frac{cov\left(d_i, d_j\right)}{\sigma_i \sigma_j}, \tag{1}$$

where cov is the covariance and σ_i and σ_j are standard deviations.

Backup files are generated and saved in the working directory. As the *DynDen* analysis is memory intensive and may take a long time, these files can be used to speed up regeneration of the data or for customized re-plotting. During execution, *DynDen* will automatically seek the existence of backup files and perform analysis only if they are not found. The convergence of a system component is characterized by an absence of change throughout time in the DTS plot and a high correlation over a long period of time in the PDC plot. The latter takes the shape of a triangular region featuring consistently high correlation. A whole system can only be considered as fully converged when all its components have converged.

3. Results and discussion

We demonstrate the usage *DynDen* on systems of increasing complexity, based upon a selection from ref. [6–8], also discussing the new insights the analysis enables.

3.1. Effect of simulation box size on a two-component system

We compare two simulations of a graphene layer in a box of ethanol (SI, Systems 1 and 3). These simulations were performed to obtain information on the interactions between exfoliating solvents with graphene. While systems are of a same chemical composition, they are very different in size: the smaller (SI, System 1) is $\sim 4 \times 4 \times 7$ nm^3 and comprises of 12 500 atoms, while the larger (SI, System 3) is $\sim 11 \times 13 \times 16$ nm^3 box and is made up of 230 000 atoms. An RMSD analysis indicates that the smaller system reaches convergence within 5 ns, while the larger system is still evolving after 40 ns. The DTS and PDC plots produced by *DynDen* feature high linear density correlation (indicated by large light coloured regions in the PDC plot) throughout the whole simulation for both systems, indicating rapid convergence. Thus, while the ethanol bulk molecules takes long time to explore their available space, the arrangement close to the graphene layer is quickly established. Therefore, in a simple two-component system, where bulk properties are not of interest, a short simulation (in the order of tens of nanoseconds) of a smaller system with (smaller bulk region) is sufficient to characterize interfacial interactions.

3.2. Appearance of surface undulations upon equilibration

We analyse a step within a multi-step dehydration protocol of amino acids intercalated within layers of hydrated layered double hydroxide (LDH), the analysis is shown in Fig. 1. This simulation is part of a study of wetting-drying cycles on early Earth [6].

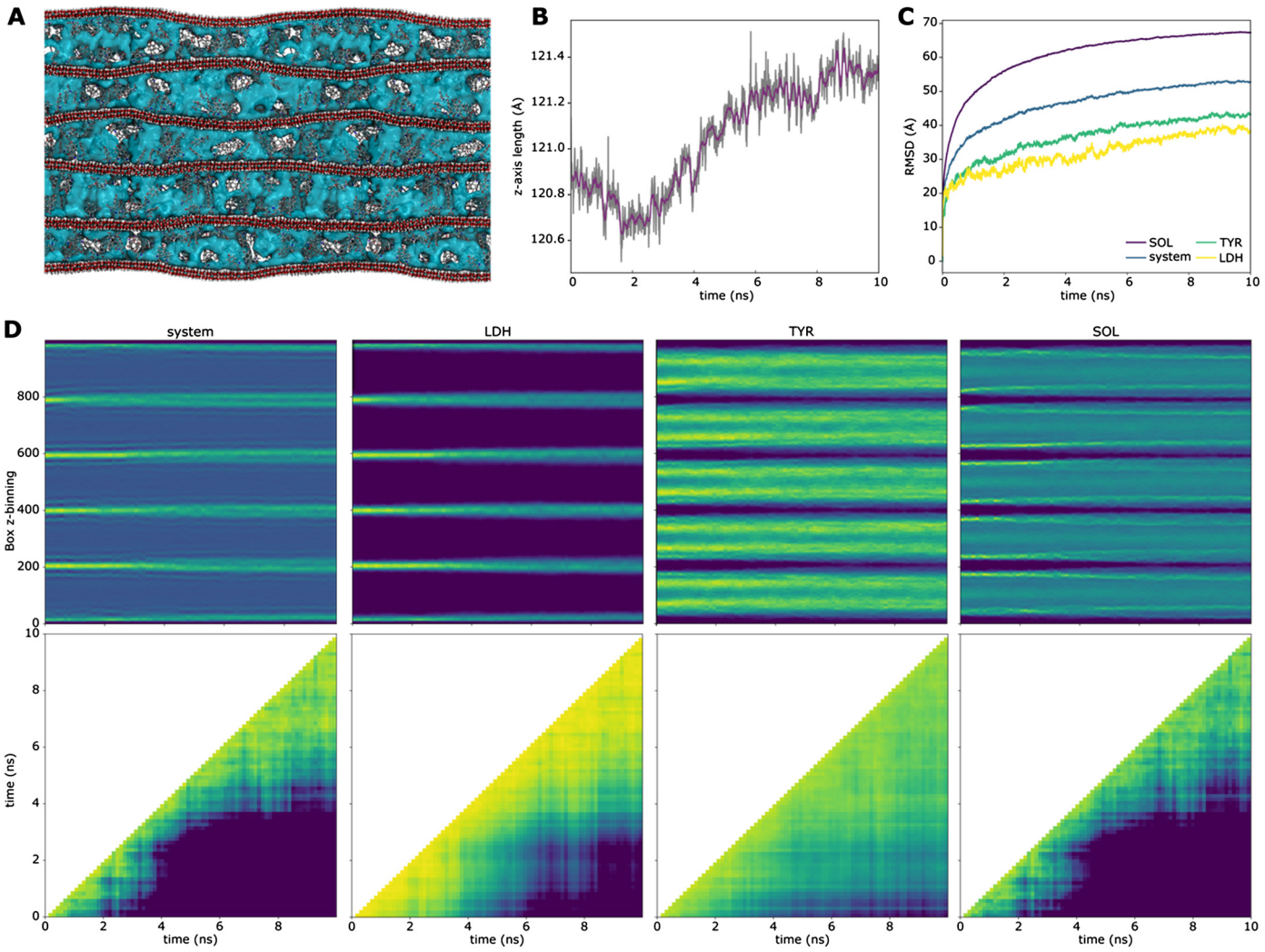

Fig. 1. Example of analysis of a system featuring tyrosine (TYR) intercalated in solvated (SOL) layered double hydroxide (LDH) from ref. [6]. **A** – System snapshot, featuring five LDH layers with interlayer tyrosine molecules and water (solvent shown as blue surface). **B** – Fluctuation of the box size in the z-direction during the 10 ns simulation. Grey line shows the size at each step, while the palatinate line is a running average. After 5 ns the box size stabilizes with fluctuation in the sub-Ångstrom scale. **C** – RMSD of the whole system and its individual components. According to this metric, the system is still equilibrating as values are still growing. **D**– Density time series (top row) and pairwise correlation coefficient plots (bottom row) for the whole system and its individual components. Light colours represent, respectively, high density and high correlation. The partial density of each system component shows high correlation for the last 5 ns of simulation, indicating that the system has converged. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

The system features five layers of LDH intercalated with tyrosine amino acids and solvated with water (see Fig. 1A). The system was first equilibrated at a higher hydration level and then, to simulate its drying, 20% of water molecules were randomly removed from the system. The simulation analysed here, starts directly from the water removal step. Within the first 2 ns the interlayer spacing decreases in order to accommodate for a reduced number of water molecules (see z-axis size on Fig. 1B). Interestingly, then the z-axis begins to expand again, becoming $\sim$1 Å bigger than at higher hydration. Upon inspection of the simulation, we notice the appearance of static undulations in the flexible LDH layers (Fig. 1A). On the DTS of the LDH (Fig. 1D) we can observe that the density peaks of individual layers smear and, after 4 ns, become bimodal. Upon visual inspection of the LDH behaviour in the simulation, we notice the appearance of static undulations in the flexible LDH layers (Fig. 1A). This behaviour is due to the bridging of the tyrosine molecules between opposing surfaces, locally pulling LDH layers together, and displacing the water into adjacent pockets (Fig. 1A). Thus, *DynDen* provided a visual representation for the observed appearance of large static undulations in the LDH, that may not be apparent by other analysis methods. The PDC of LDH (Fig. 1D) features a prominent light coloured area (Pearson correlation coefficient equal to 1) which, therefore, reveals that individual partial densities remain unchanged for the last 5 ns of simulation. This is indicative that static undulations identified by the *DynDen* method are a long-lived structural feature.

3.3. Appearance of slow transient dynamics in multi-component systems

We examine the converge of a complex system, featuring a kaolinite clay intercalated with decanoic acid and brine solution with Na^+, Ca^{2+} and Cl^- ions (SI, System 5). This model is designed to simulate the process of low salinity enhanced oil recovery, where a brine is injected into a reservoir to facilitate the removal of oil components [8]. Kaolinite is a clay mineral common in oil reservoirs featuring both a hydrophilic aluminol and a hydrophobic siloxane surfaces. From RMSD analysis it remains unclear whether the whole system has reached convergence. Meanwhile, DTS analysis of components shows that within 20 ns decanoic acid rearranges from its original surface-coating starting configuration. First, the decanoic acid fully desorbs from the kaoli-

nite surfaces. It then forms a droplet on the hydrophobic surface, and reabsorbs onto the hydrophilic surface, forming multiple solvation layers. Ions show different behaviours, whereby Na^+ adsorbs on both surfaces, Cl^- adsorbs on the hydrophilic surface and Ca^{2+} surrounds the decanoic acid droplet and follows its movement throughout the simulation. Inspection of the DTS plot reveals that local convergence is quickly reached close to the hydrophilic surface, where the decanoic acid was re-adsorbed. Meanwhile, the droplet that has detached from the hydrophobic surface remains mobile, interacting transiently with both the hydrophobic surface and the oil layer adsorbed on the hydrophilic surface. These dynamics are revealed in the PDC plot as a checkerboard pattern, alternating high convergence (light colour) and low convergence (dark colour) areas. Such dynamic phenomena, where one component remains mobile, are also observable in a simulation featuring LDHs intercalated with three random 24-amino acid-long peptides (SI, System 6) [6]. The DTS plot shows that the three peptides independently adsorb and desorb from the surfaces throughout the simulation. While convergence in the PDC plot is a sufficient criterion for the identification of equilibration, for a dynamic system the DTS plot can also act as a diagnostic tool helping the visualization of the dynamics of all system components. In situations where emergence of such dynamic properties are observed, multiple independent repeats are of great value to obtain sufficient statistics for conclusive results.

3.4. Understanding slowly evolving multi-component systems

We analyse two simulations representing a realistic mixture of crude oil components interacting with clay under natural reservoir conditions [7]. The system features kaolinite clay and a mixture of methane, hexane, dodecane, octadecane, naphatalene and octadecanoic acid. The identical system was modelled at reservoir pressure of 100 bar and temperatures of 298 K and 323 K (SI, System 7a and 7b respectively). At lower temperature, the RMSD indicates that convergence is reached within 50 ns. Meanwhile, the box size is still reducing over the following 25 ns indicating that the system is still undergoing changes. This is in agreement with the PDC plot, showing that long-time correlations emerge only after 100 ns. The DTS plot of each individual component shows the formation of ordered layers spanning the whole interlayer. These layers remain stable for the remainder of the simulation. While this may indicate equilibration, chemical intuition suggests the system is more likely to be kinetically trapped. This is a plausible assumption, as the simulated temperature and pressure are close to crystallization conditions of several of the oil components, and the hydrophobic kaolinite surface further promotes co-alignment similar to paraffin crystals. When we model exactly the same system at slightly warmer reservoir temperature of 323 K, we no longer see the formation of these highly ordered layers spanning across the whole interlayer. The RMSD of individual components is not stable, and the box size continues changing until 150 ns. The PDC plots show that the system continues to evolve, with substantial changes occurring at 50 and 125 ns. The DTS plot reveals that changes are associated with the rearrangement of components onto the hydrophobic surface, where molecules are adsorbed along the clay surface. Due to strong interactions between surface and oil components, these interactions are long-lived. Overall, when modelling such complex systems at extreme conditions, the scientist should ask themself whether this is truly representative of the experimental studies, and whether limits of free molecular dynamics have been reached and enhanced sampling methods such as replica exchange should be used.

4. Conclusion

We have presented a method and associated software, *DynDen*, aiding the assessment of evolution in molecular systems featuring interfaces. The software provides a quantitive way of evaluating convergence of simulations. Additionally, *DynDen* proves itself as a useful tool for identifying slow dynamical processes occurring in long simulations, which can be easily missed with conventional analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Shan Jian and Tom Underwood for having shared their raw simulation data with us and Durham University HPC facility, Hamilton, for computer time. VE would like to thank the University of Edinburgh for the award of the Chancellor's Fellowship. MTD acknowledges support of an EPSRC fellowship (EP/P016499/1).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cpc.2021.108126.

References

[1] K. El Hage, F. Hédin, P.K. Gupta, M. Meuwly, M. Karplus, eLife 7 (2018) e35560, https://doi.org/10.7554/eLife.35560.
[2] V. Gapsys, B.L. de Groot, eLife 8 (2019) e44718, https://doi.org/10.7554/elife.44718.
[3] V. Gapsys, B.L. de Groot, eLife 9 (2020) 1–21, https://doi.org/10.7554/eLife.57589.
[4] B. Knapp, L. Ospina, C.M. Deane, J. Chem. Theory Comput. 14 (2018) 6127–6138, https://doi.org/10.1021/acs.jctc.8b00391.
[5] V.N. Maiorov, G.M. Crippen, Proteins, Struct. Funct. Bioinform. 22 (3) (1995) 273–283, https://doi.org/10.1002/prot.340220308.
[6] V. Erastova, M.T. Degiacomi, D.G. Fraser, H.C. Greenwell, Nat. Commun. 8 (2017), https://doi.org/10.1038/s41467-017-02248-y.
[7] S. Tian, et al., Energy Fuels 32 (2018) 1155–1165, https://doi.org/10.1021/acs.energyfuels.7b02763.
[8] T. Underwood, V. Erastova, H.C. Greenwell, J. Phys. Chem. C 120 (2016) 11433–11449, https://doi.org/10.1021/acs.jpcc.6b00187.
[9] T. Underwood, V. Erastova, P. Cubillas, H.C. Greenwell, J. Phys. Chem. C 119 (2015) 7282–7294, https://doi.org/10.1021/acs.jpcc.5b00555.
[10] R. Gowers, et al., in: Proceedings of the 15th Python in Science Conference, 2016, https://doi.org/10.25080/majora-629e541a-00e.
[11] N. Michaud-Agrawal, E.J. Denning, T.B. Woolf, O. Beckstein, J. Comput. Chem. 32 (10) (2011) 2319–2327, https://doi.org/10.1002/jcc.21787.